

Estratégias de previsão fiscal: um estudo empírico para a economia brasileira

Rennan Kertlly de Medeiros*

Edilean Kleber da Silva Bejarano Aragón †

Cássio da Nóbrega Besarria ‡

Resumo

Este artigo objetiva realizar a previsão da arrecadação federal do Brasil a partir de diversos métodos de aprendizado de máquina, para diferentes amostras. Para previsão da variável de interesse foram utilizadas 34 variáveis explicativas. A metodologia de estimação utilizada está dividida em três categorias de aprendizagem: shrinkage, ponderação e fator. Os resultados sugerem que o modelo Elastic Net possui a maior acurácia para previsões mensais com uso de 20% e 30% da amostra para teste, para períodos curtos. Para previsão de períodos acumulados, o modelo LASSO tem alta performance. Por fim, constatamos que o modelo Bagging é limitado com aumento de períodos de previsão e, sobretudo, com diminuição de frequência das variáveis.

Palavras-chave: Previsão fiscal; Arrecadação; Modelos de machine learning; Elastic Net.

Abstract

This article aims to forecasting Brazil's federal tax revenues from different machine learning methods, for different samples. To forecasting the variable of interest, 34 explanatory variables were used. The estimation methodology used is divided into three learning categories: shrinkage, ensemble and factor. The results suggest that the Elastic Net model has the highest accuracy for monthly forecasts using 20% and 30% of the test sample, for short periods. For forecasts of accumulated periods, the LASSO model has high performance. Finally, we verify that the Bagging model is limited with an increase in forecast periods and, above all, with a decrease in the frequency of variables.

Keywords: Fiscal forecasting; Tax revenues; Machine learning models; Elastic Net.

JEL Code: C53; E17; H27.

Área 4: Macroeconomia, Economia Monetária e Finanças

*Doutorando em Economia pelo PPGE-UFPB. E-mail: rennanmedeiros22@hotmail.com.

†Professor do PPGE-UFPB. E-mail: edilean@hotmail.com.

‡Professor do PPGE-UFPB. E-mail: cassiodanobrega@yahoo.com.br.

1 Introdução

O período de pandemia da Covid-19 trouxe um cenário de incerteza para a economia global, onde os países mais afetados são aqueles em processo de desenvolvimento. Por exemplo, o Brasil enfrenta dificuldades para retomada do crescimento pré-pandemia, acompanhada de elevados gastos governamentais com política de transferência de renda. Nesse sentido, a antecipação do comportamento de variáveis fiscais consiste em uma estratégia para formulação de políticas governamentais, com responsabilidade fiscal.

De modo geral, a literatura internacional para previsão sobre variáveis macroeconômicas é ampla, com estudos voltados para a atividade econômica, e também estratégias de previsão para variáveis fiscais. Com o uso de modelo econométrico univariado, ARIMA, Kamlet, Mowery e Su (1987) comparam suas previsões de curto e longo prazos com as previsões elaboradas pelo Poder Executivo (OMB, sigla do órgão em inglês) e pelo Congresso (CBO, sigla em inglês) norte-americano, para taxa de crescimento real da economia, inflação e desemprego. Os autores indicam que para o período de um ano (curto prazo), as previsões são semelhantes. Por sua vez, Plesko (1988) avalia as previsões do CBO e OMB para o PIB nominal, despesas correntes, receitas correntes e do déficit, e indicam resultados parecidos para as previsões de curto prazo. Enquanto, Krause e Douglas (2005) avaliam que as previsões do CBO e OMB para PIB real, inflação, desemprego, receitas do setor público, despesas fiscais e déficit orçamentário, de 1976 a 2001, são estatisticamente iguais, com exceção das variáveis desemprego e receitas. Outros estudos também comparam projeções macroeconômicas, por exemplo, Frankel (2011) e Kliesen, Thornton et al. (2012)

Para a economia brasileira, alguns estudos realizam previsões para variáveis macroeconômicas (por exemplo, Zuanazzi e Ziegelmann (2014), Caldeira, Moura e Santos (2015), Medeiros e Mendes (2016) e Garcia, Medeiros e Vasconcelos (2017)), entretanto, abordagens para variáveis fiscais ainda são limitadas. Nesse contexto, Campos (2009) a partir de diversos modelos econométricos univariados e multivariados, desenvolve previsões para a arrecadação tributária brasileira, em virtude da importância dessa variável para construção do orçamento anual do governo brasileiro. O estudo indica a relevância do uso de métodos de séries temporais no desempenho da previsão da arrecadação tributária, quando comparados com método dos indicadores, até então utilizado pela Receita Federal Brasileira (RFB).

Além de poucos estudos sobre previsão de variáveis macroeconômicas brasileiras, nesse caso variáveis fiscais, os métodos utilizados consideram poucas variáveis preditivas. Nesse sentido, este estudo adota e compara inúmeros algoritmos de aprendizado de máquina, em que estabelecemos o melhor modelo econométrico para previsão fiscal. Em um cenário macroeconômico diferente e para a economia norte-americana, Gu, Kelly e Xiu (2020) indicam que métodos de *machine learning*, juntamente com grande número de preditores, melhoram a acurácia de previsão. Por sua vez, Medeiros et al. (2021) indicam a superioridade de modelos da classe *machine learning* quando comparados com modelos univariados simples, por exemplo, passeio aleatório, arima, dentre outros.

Diante da importância do cenário fiscal após a pandemia de Covid-19 e escassez de estudos voltados para esse contexto, este estudo objetiva testar e comparar modelos de aprendizado de máquina na busca por um modelo de maior acurácia para previsão da principal variável orçamentária, a arrecadação tributária federal (preliminar na formação da responsabilidade fiscal do governo federal).¹

¹ Ver Lei do Orçamento Anual (LOA) e Lei de Responsabilidade Fiscal (Lei Complementar nº 101, de 04/05/2000).

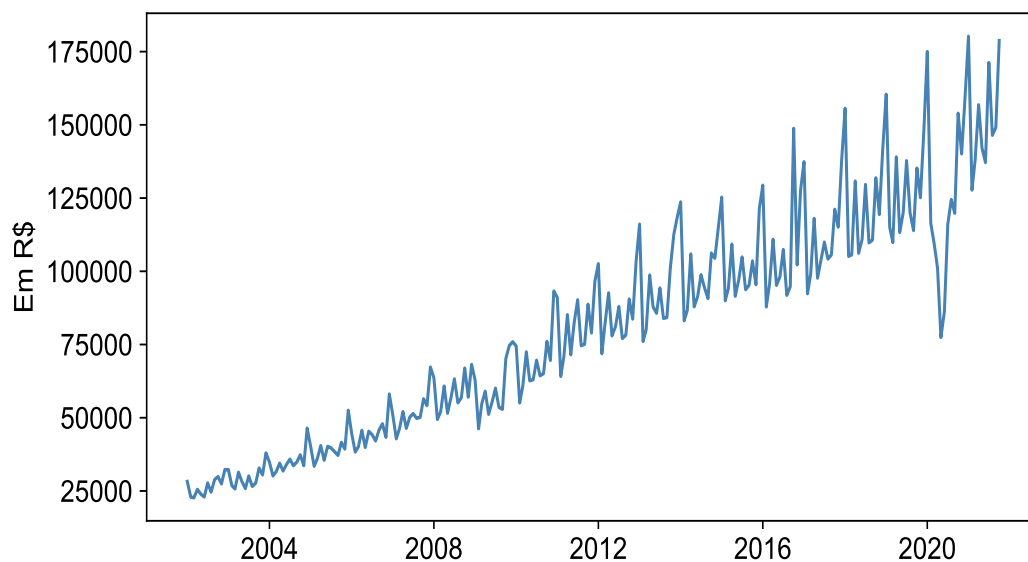
Este estudo contribui com a literatura de previsão fiscal em três sentidos: i) como modelos consolidados na literatura de previsão, estimamos e comparamos a acurácia de previsão dos modelos de aprendizado de máquina; ii) realizamos a previsão da arrecadação tributária mensal brasileira para um longo período amostral (2002 a 2021), algo jamais realizado para a economia brasileira; por fim iii) realizamos um exercício econométrico robusto para frequência trimestral, essencial para tomadas de decisão de gestores públicos. Portanto, o resultado dessa pesquisa pode auxiliar na formulação de políticas governamentais para antecipação e controle do orçamento público.

Além desta introdução, este estudo está estruturado em mais três seções. A Seção 2 é destinada para apresentação do conjunto de dados utilizados na previsão da variável fiscal. A Seção 3 apresenta a metodologia executada. A Seção 4 é destinada para os resultados e discussão da pesquisa. O exercício de robustez está na Seção 5. Por fim, a Seção 6 contém as considerações finais do estudo.

2 Banco de Dados

Esta seção é reservada para apresentação do conjunto de dados utilizados para previsão da variável fiscal de interesse, arrecadação tributária federal. O período amostral utilizado na previsão varia de janeiro de 2002 a outubro de 2021, com frequência mensal, o que totaliza 238 observações. O período amostral é justificado pela limitação de observações para todas as variáveis usadas na estimação, ou seja, a arrecadação e as variáveis preditoras. A Figura 1 ilustra a trajetória da arrecadação tributária ao longo do tempo.

Figura 1 – Trajetória da arrecadação fiscal (Em Milhões de R\$)



Fonte: Elaboração dos autores.

De acordo com a trajetória da arrecadação, observamos uma forte oscilação de seus valores ao longo dos anos. Entretanto, em meados de 2020 houve uma queda brusca na arrecadação brasileira, o que pode ser justificado pelo início da pandemia de Covid-19. Em virtude da escolha de variáveis preditivas para a arrecadação, optamos por modelos de previsão de aprendizado de máquina, o que requer um grande número de preditores.

Para este estudo, além da variável de interesse (arrecadação), utilizamos 34 variáveis econômicas que possivelmente explicam o comportamento da arrecadação tributária. Em virtude do grande número de preditores, a técnica de *machine learning* sobressai na escolha, já que possui algoritmos de penalização com intuito de seleção de variáveis preditoras.

Tabela 1 – Resumo das variáveis utilizadas na predição de arrecadação

Grupo	Fonte	Código
Atividade econômica		
Papelão ondulado	IBGE	-
Produção industrial (produtos de borracha e material plástico)	IBGE	-
Produção industrial (celulose, papel e produtos de papel)	IBGE	-
Produção (geral)	Banco Central	21859
Produção (indústria de transformação)	Banco Central	21862
Produção (bens de capital)	Banco Central	21863
Produção (bens intermediários)	Banco Central	21864
Produção (bens de consumo)	Banco Central	21865
Produção total de autoveículos	Banco Central	1373
Vendas de autoveículos (total)	Banco Central	1378
Consumo de energia elétrica (comercial)	Banco Central	1402
Consumo de energia elétrica (industrial)	Banco Central	1404
Consumo de energia elétrica (total)	Banco Central	1406
Volume vendas no varejo (automóveis, motocicletas, partes e peças)	Banco Central	1548
Índice Geral de Preços do Mercado (IGP-M) - Var. % mensal	Banco Central	189
Índice Geral de Preços - disponibilidades Interna (IPA-M) - Variação % mensal	Banco Central	190
Índice de Preços por Atacado-Mercado (IPA-M) - Variação % mensal	Banco Central	7450
Índice Nacional de Preços ao Consumidor - Amplo (INPC) - Itens livre - Var % mensal	Banco Central	188
Índice Nacional de Preços ao Consumidor Amplo (IPCA) - Itens livres - Variação % mensal	Banco Central	11428
Produção de caminhões	Banco Central	1375
Índice Volume de Vendas no Varejo	Banco Central	1455
Finanças públicas		
Contribuição - Finsocial / Cofins - Total - Receita bruta	Ministério da Economia \SRF	-
Imposto sobre a Renda - Pessoas Físicas	Ministério da Economia \SRF	-
Imposto sobre a Renda - Pessoas Jurídicas	Ministério da Economia \SRF	-
Dívida Líquida do Setor Público (% PIB) - Interna - Setor Público Consolidado	Banco Central	4524
Dívida Líquida do Setor Público (% PIB) - Total - Setor Público Consolidado	Banco Central	4513
Dívida Líquida do Setor Público (% PIB) - Externa - Setor Público Consolidado	Banco Central	4535
Dívida Líquida do Governo Geral (% PIB)	Banco Central	4536
Dívida Fiscal Líquida - Saldos (% PIB)	Banco Central	10831
NFSP sem desvalorização cambial (% PIB)	Banco Central	5717
Dívida mobiliária (saldos)	Banco Central	2213
Indicador de crédito		
Saldo da carteira de crédito - Total	Banco Central	20539
Mercado financeiro e de capitais		
Taxa de juros - Selic	Banco Central	4390
Setor Externo		
Taxa de câmbio	Banco Central	3697

Fonte: Elaboração dos autores.

Baseado na técnica de [McCracken e Ng \(2016\)](#), classificamos as variáveis explicativas em cinco grupos, são eles: atividade econômica, finanças públicas, indicador de crédito, mercado financeiro e de capitais e setor externo.

Para o grupo de Atividade Econômica, foram utilizadas 21 variáveis. As variáveis papelão ondulado, produção industrial (produtos de borracha e material plástico) e produção industrial (celulose, papel e produtos de papel) foram extraídas do Instituto Brasileiro de Geografia e Estatística (IBGE). As demais 18 variáveis ligadas à atividade econômica foram coletadas no Banco Central do Brasil, são elas: produção geral, produção (indústria de transformação), produção de bens de capital, produção de bens intermediários, produção de bens de consumo, produção total de autoveículos, vendas totais de autoveículos, consumo de energia elétrica (comercial), consumo de energia elétrica (industrial), consumo de energia elétrica (total), volume de vendas no varejo (automóveis, motocicletas, partes e peças), Índice Geral de Preços do Mercado (IGP-M), Índice Geral de Preços - disponibilidades internas (IPA-M), Índice de Preços por Atacado-Mercado (IPA-M), Índice Nacional de Preços ao Consumidor (INPC), Índice Nacional de Preços ao Consumidor Amplo, itens livres, produção de caminhões e índice de volume de vendas no varejo.

O grupo Finanças Públicas é composto por dez indicadores. As variáveis Contribuição - Finsocial/Cofins, imposto sobre a renda (pessoas físicas) e imposto sobre a renda (pessoas jurídicas) foram extraídas do Ministério da Economia/Secretaria da Receita Federal. As demais sete variáveis foram extraídas do Banco Central, são elas: Dívida Líquida do Setor Público (DLSP) interna do setor público consolidado, DLSP total do setor público consolidado, DSLP externa do setor público consolidado, dívida líquida do governo geral, saldos da dívida fiscal líquida, Necessidade de Financiamento do Setor Público (NFSP) sem desvalorização cambial e dívida imobiliária.

As três variáveis restantes foram coletadas no banco central e são subdivididas em três grupos restantes. O grupo Indicador de Crédito considera a variável saldo da carteira de crédito total; o grupo Mercado Financeiro e de Capitais inclui a taxa de juros básica da economia brasileira (SELIC); por fim, o grupo Setor Externo considera a taxa de câmbio comercial (compra). A Tabela 1 resume o conjunto de variáveis usadas como possíveis preditoras da arrecadação tributária do governo federal. As estatísticas descritivas das variáveis preditoras e da arrecadação (variável de interesse) estão descritas no Apêndice A.

3 Metodologia e Modelos de aprendizado de máquina

Esta seção é reservada para apresentação da metodologia utilizada para previsão de arrecadação tributária, bem como são apresentadas especificações de nove modelos de *machine learning* utilizados.

3.1 Metodologia

Para estimação dos inúmeros modelos de aprendizado de máquina, seguimos o seguinte modelo:

$$T_{t+h} = G_h(x_t) + u_{t+h}, \quad h = 1, \dots, H, \quad t = 1, \dots, T \quad (1)$$

em que T_{t+h} é a arrecadação no mês $t + h$; $x_t = (x_{1t}, \dots, x_{nt})'$ é um vetor n de co-variáveis que indicam prováveis defasagens de T_t e/ou fatores comuns, bem como um conjunto de

preditores candidatos; $G(\cdot)$ é o mapeamento entre co-variáveis e arrecadação futura; e u_t é um erro aleatório de média zero. A função alvo $G_h(\cdot)$ pode ser um único modelo ou um conjunto de diferentes especificações. Os mapeamentos são diferentes conforme varia o horizonte de previsão.

A equação de previsão objetiva é dada por:

$$\hat{T}_{t+h|t} = \hat{G}_{h,t-R_h+1:t}(x_t) \quad (2)$$

em que $\hat{G}_{h,t-R_h+1:t}$ é uma função objetivo estimada baseada em uma série temporal $t - R_h + 1$ até t e R_h é o tamanho da janela, que varia conforme o horizonte de previsão e o número de variáveis defasadas no modelo. Seguindo [Medeiros et al. \(2021\)](#), consideramos as previsões diretas já que não realizamos nenhuma tentativa de previsão das co-variáveis.

As previsões realizadas para a variável de interesse, arrecadação, será baseada em uma estrutura de janela móvel com tamanho determinado. Em outras palavras, para o período amostral de janeiro de 2002 até outubro de 2021, o número de observações é $R_h = 238 - h - p - 1$, em que p é o número de defasagens do modelo. A escolha de previsão por janela móvel é justificada a partir de dois pontos: i) diminuir os efeitos de possíveis quebras estruturais e *outliers*; e ii) evitar problemas de execução de testes de desempenho preditivos superiores entre modelos econométricos aninhados.²

Dentre os modelos de aprendizado de máquina, consideramos os métodos de *shrinkage* que estabelecem uma seleção de variáveis com objetivo de elevar a performance preditiva da variável alvo ou de interesse. O processo de estimação envolve a penalização de estimadores (preditores) irrelevantes como penalização. É comum os coeficientes estimados dos preditores (variáveis explicativas) irrelevantes serem reduzidos na previsão, em que podem tender a zero.

A partir de redução dos valores absolutos dos parâmetros, os modelos *shrinkage* selecionam variáveis com maior capacidade de predição da variável alvo, mecanismo conhecido como regularização. De modo geral, métodos de *shrinkage* executam o processo de penalização abaixo:

$$\min_b ||y_t - z_t' b||^2$$

sujeito a $g(b) \leq s$.

em que $||y_t - z_t' b||^2$ representa a soma do quadrado dos resíduos, s denota um parâmetro de ajuste de penalização e $g(b)$ é uma função penalizadora dos coeficientes b da expressão.

Neste estudo, utilizamos cinco estimadores *shrinkage*, são eles: Lasso, Elastic-Net, Ridge, Lasso Adaptativo (LassoAdap) e Elastic-Net Adaptativo (LassoAdap). Também utilizamos modelos que envolve métodos de média em conjunto, nesse estudo optamos por Bagging e CSR. Além desses modelos, também usamos dois métodos baseados na extração por Fatores, nesse caso optamos pelo Fatores *Target*, baseado em [Bai e Ng \(2008\)](#), e o Boosting, baseado em [Bai e Ng \(2009\)](#). Além dos modelos destacados, também analisamos dois modelos alternativos: modelo com média simples e outro baseado na mediana, de todos os modelos de aprendizado estatísticos utilizados na predição da variável alvo. Os horizontes de previsões para todos os modelos especificados são para oito períodos à frente, previsão acumulada de quatro períodos à frente e, por fim, previsão acumulada com oito períodos à frente.

² Ver [Giacomini e White \(2006\)](#).

3.2 Shrinkage

Para estimadores *shrinkage* de modelos lineares executados neste estudo temos que:

$$\hat{\beta}_h = \arg \min_{\beta_h} \left[\sum_{t=1}^{T-h} (y_{t+h} - \beta_h' x_t)^2 + \sum_{i=1}^n p(\beta_{h,i}; \lambda, \omega_i) \right] \quad (3)$$

em que $p(\beta_{h,i}; \lambda, \omega_i)$ é uma função de penalidade que depende do parâmetro de penalidade λ e em um peso $\omega_i > 0$. Consideramos tipos de escolhas diferentes para a função de penalidade.

3.2.1 Ridge

A Regressão Ridge foi proposta por [Hoerl e Kennard \(1970\)](#). A função de penalidade é dada por:

$$\sum_{i=1}^n p(\beta_{h,i}; \lambda, \omega_i) := \lambda \sum_{i=1}^n \beta_{h,i}^2 \quad (4)$$

A principal vantagem dessa regressão é a facilidade de cálculo de sua solução analítica, bem como reduz coeficientes ligados às variáveis menos relevantes na predição, para quase zero. Entretanto, os coeficientes quase nunca atinge zero para qualquer tamanho de λ .

3.2.2 LASSO

Este método foi proposto por [Tibshirani \(1996\)](#), em que a função de penalidade é dada por:

$$\sum_{i=1}^n p(\beta_{h,i}; \lambda, \omega_i) := \lambda \sum_{i=1}^n |\beta_{h,i}| \quad (5)$$

Este operador reduz as variáveis operadoras para zero, eliminando-as da estimação quando forem irrelevantes. Todavia, a consistência de seleção do modelo é obtida apenas sob hipóteses rigorosas. Quando seleciona as variáveis, o LASSO minimiza a soma dos quadrados dos resíduos, desde que a soma do valor absoluto dos coeficientes seja menor que uma constante. Além disso, com $\lambda = 0$ a estimação será por mínimos quadrados ordinários (OLS).

Conforme [Zou e Hastie \(2005\)](#), apesar de renomado para estudos de previsão, o LASSO apresenta algumas limitações em algumas situações. Aqueles autores listam alguns exemplos: i) quando o número de preditores é superior ao número de observações, aquele operador seleciona apenas a quantidade de preditores equivalente ao número de observações, em virtude do problema de otimização convexa inerente ao método; no caso de efeito de agrupamento, o LASSO quase sempre seleciona apenas uma variável por grupo; e, por fim, quando o número de observações é superior ao número de preditores e na presença de fortes correlações entre os preditores, é esperado que a regressão Ridge tenha performance superior ao LASSO.

3.2.3 LASSO Adaptativo

Este método foi proposto por [Zou \(2006\)](#) para melhora de desempenho do modelo LASSO convencional. Este método adaptativo utiliza a mesma penalidade do LASSO, com a inclusão de um parâmetro de ponderação. Agora, a função de penalidade é dada por:

$$\sum_{i=1}^n p(\beta_{h,i}; \lambda, \omega_i) := \lambda \sum_{i=1}^n \omega |\beta_{h,i}| \quad (6)$$

em que $\omega = |\beta_{h,i}^*|^{-1}$ e $\beta_{h,i}^*$ é o coeficiente do primeiro passo da estimação LASSO. O LASSO Adaptativo suporta mais variáveis do que observações e possui performance alta em processos não gaussianos e sob heteroscedasticidade.³

3.2.4 Elastic Net

Este método foi proposto por [Zou e Hastie \(2005\)](#). Elastic Net é uma generalização que incluem casos especiais como LASSO e Ridge. A função de penalidade é definida por:

$$\sum_{i=1}^n p(\beta_{h,i}; \lambda, \omega_i) := \alpha \lambda \sum_{i=1}^n \beta_{h,i}^2 + (1 - \alpha) \lambda \sum_{i=1}^n |\beta_{h,i}| \quad (7)$$

em que $\alpha \in [0, 1]$. Semelhante ao LASSO, Elastic Net realiza de forma simultânea seleção automática e encolhimento contínuo, em que é possível a seleção de grupos de variáveis correlacionadas.

Elastic Net é uma combinação convexa das normas ℓ_1 e ℓ_2 . Em outras palavras, aquele método é vantajoso porque incorpora dois tipos de penalidade específicas: penalidade normada de ordem 2, estabelecida por $\sum_{i=1}^n \beta_{h,i}^2$, e penalidade normada de ordem 1, estabelecida por $\sum_{i=1}^n |\beta_{h,i}|$. No primeiro caso, se $\alpha = 1$, o método se reduz ao Ridge. Por outro lado, no segundo caso, com $\alpha = 0$, temos como resultado o operador proposto por [Tibshirani \(1996\)](#), ou seja, o LASSO.

A obtenção dos valores de α é através de validação cruzada. Entretanto, este método não é indicado para séries temporais em virtude da correlação entre as variáveis preditoras. Nesse sentido, [Bai e Ng \(2008\)](#) sugerem a utilização de critérios de escolha alternativos, como Critério de Informação de Akaike (AIC) e Critério Bayesiano de Schwarz (BIC), para obtenção dos parâmetros de ajuste. Nesse estudo, optamos pelo critério BIC.

Seguindo [Kock e Callot \(2015\)](#) e [Medeiros e Mendes \(2016\)](#), para todos métodos da classe LASSO, o parâmetro de penalização λ é escolhido a partir do critério BIC. O parâmetro α para a penalização Elastic Net é definida em 0,5. Os pesos das extensões adaptativas de LASSO e Elastic Net são definidas por $\omega_i = \frac{1}{|\tilde{\beta}_i| + \frac{1}{\sqrt{T}}}$, em que $\tilde{\beta}_i$ é o estimador do método adaptativo.

3.3 Métodos de ponderação (*ensemble*)

As previsões de conjunto são resultados de uma ponderação das previsões de vetor de métodos.

³ Ver [Medeiros e Mendes \(2016\)](#).

3.3.1 Bagging

O método Bagging foi proposto por [Breiman \(1996\)](#). Aquele método combina previsões de um conjunto de modelos instáveis estimados para diferentes amostras de *bootstrap*, ou seja, consiste de um método de agregação através de re-amostragem. Quase sempre a combinação é melhorada quando os modelos envolvidos são muito diferentes. As etapas do Bagging são:

Passo 1: para cada amostra do *bootstrap*, executa uma regressão OLS com as variáveis candidatas e seleciona aquelas com estatística t absoluta superior a um determinado limite c .

Passo 2: estima uma nova regressão apenas com variáveis selecionadas no passo 1.

Passo 3: os coeficientes da segunda regressão (passo 2) são usados para computar as previsões da amostra atual.

Passo 4: As três primeiras etapas para amostras de *bootstrap* B são repetidas e a previsão final como a média das previsões de B é calculada.

Em termos matemáticos, temos:

$$y_{t+h}^{bagg} = W_t \hat{\beta}^{OLS} + \sum_{j=1}^r \Gamma(t_j) \hat{\beta}_{F_j} \hat{F}_j \quad (8)$$

em que $\hat{\beta}^{OLS}$ é a estimação de OLS de y_{t+h} contra W_t ; $\hat{\beta}_{F_j}$ é a regressão OLS residual; $j = 1 \dots r$ estabelece a quantidade de fatores estimados; e t_j é a estatística *t-student* associada a $\hat{\beta}_{F_j}$.

$$\Gamma(t_j) = 1 - \psi(t+c) + t - 1[\phi(t-c) - \phi(t+c)] \quad (9)$$

em que ψ é a densidade normal padrão e ϕ representa uma função de distribuição acumulada da distribuição normal.

O número de replicações para o método Bagging é $B = 100$. O procedimento de pré-teste adotado é baseado em [Inoue e Kilian \(2008\)](#), em que o nível estabelecido é de 5%.

3.3.2 CSR

O método de Regressões de Subconjunto Completadas (CSR) foi proposto por [Elliott, Gargano e Timmermann \(2013\)](#) e [Elliott, Gargano e Timmermann \(2015\)](#). A seleção de subconjunto ótimo de x_t para previsão de T_{t+h} com testagem de todas as combinações de regressores é computacionalmente muito exigente e, quase sempre, inviável. Suponha que temos n variáveis candidatas. O objetivo é a seleção de um número $q \leq n$ variáveis e execução de regressões com uso de todas as combinações possíveis de q de n variáveis.

O CSR suporta um número pequeno de variáveis. Para grandes conjuntos de variáveis, o número de regressões a serem estimadas aumenta rápido. Dessa forma, iniciamos a regressão A_{t+h} em cada uma das possíveis variáveis candidatas (incluindo defasagens) e guardamos as estatísticas t de cada variável. As estatísticas t são classificadas por valor absoluto, e selecionamos as \tilde{n} variáveis mais relevantes na classificação. A previsão CSR é calculada sobre essas variáveis. Semelhante ao teste Bagging, o pré-teste executado é ao nível de 5%.

3.4 Modelos de Fatores

A literatura sobre modelos de fatores é bem representada no estudo de Bai (2003). O modelo fatorial reduz a dimensão do modelo completo a partir de extração de componentes comuns de todos preditores. Os fatores são calculados como componentes principais de um grande conjunto de variáveis z_t tal que $F_t = Az_t$, em que A é uma matriz de rotação e F_t representa o vetor de componentes principais. De acordo com a Equação 1, temos que x_t é dado por T_{t+h} , $j = 0, 1, 2, 3$ mais f_{t-j} , $j = 0, 1, 2, 3$, em que f_t é o vetor com os quatro primeiros componentes principais de z_t .

Para esse estudo, o número de fatores e defasagens dos modelos é definido em quatro. Para os fatores *target*, o nível de significância de 5% ($\alpha = 0.05$). O número máximo de iterações estabelecido é $M = 10 \times$ número de variáveis. Entretanto, o algoritmo será parado quando o critério de escolha BIC começa aumentar. Seguindo Bai e Ng (2008) e Medeiros et al. (2021), o parâmetro *shrinkage* é definido por $v = 0, 2$.

3.4.1 Fatores Target

Bai e Ng (2008) sugerem este procedimento para aumento da performance das previsões por fatores. A proposta dessa metodologia por fatores direcionados é o cálculo dos componentes principais somente das variáveis com alto poder de predição da arrecadação futura. Seja $z_{i,t}$, $i = 1, \dots, q$ variáveis candidatas e w_t um conjunto de controles. Conforme Bai e Ng (2008) e Medeiros et al. (2021), usamos os valores defasados da variável de interesse, nesse caso a arrecadação, como controles. O procedimento ocorre em quatro etapas:

Passo 1: para $i = 1, \dots, q$, regrida T_{t+h} em w_t e $z_{i,t}$ e calcule as estatísticas t para os coeficientes correspondentes em $z_{i,t}$.

Passo 2: Escolha um nível de significância α e selecione as variáveis que são significativas com uso das estatísticas t calculadas.

Passo 3: Deixe $z_t(\alpha)$ ser as variáveis selecionadas nos passos 1 e 2. Estime os fatores F_t de $z_t(\alpha)$ pelos componentes principais.

Passo 4: Regrida T_{t+h} em w_t e f_{t-j} , $j = 0, 1, 2, 3$, em que $f_t \subset F_t$. Para esse caso, o número de fatores em f_t é selecionado a partir do critério de informação de escolha BIC.

3.4.2 Fatores Boosting

Seguindo Bai e Ng (2008), deixe $z_t \in \mathbb{R}$, o conjunto dos n fatores calculados a partir das n variáveis originais mais quatro defasagens de cada fator. Dessa forma, $q = 5n$. O algoritmo é definido a partir de três passos, a saber:

Passo 1: Deixe $\Phi_{t,0} = \bar{T}$ para cada t , em que $\bar{T} = \frac{1}{t} \sum_{i=1}^t T_i$.

Passo 2: Para $m = 1, \dots, M$: a) Calcule $\hat{u}_t = T_t - \Phi_{t-h,m-1}$. b) Para cada variável candidata $i = 1, \dots, q$, regrida o resíduo atual em $z_{i,t}$ para obter \hat{b}_i , e compute $\hat{e}_{t,i} = \hat{u}_t - z_{i,t}\hat{b}_i$. Calcule $SSR_i = \hat{e}'_i \hat{e}_i$. c) Selecione i_m^* como índice da variável que fornece menor SSR e defina $\hat{\phi}_{t,m} = \hat{\phi}_{t,m-1} + v\phi_{t,m}$, em que v é o tamanho da etapa.

Passo 3: Pare o algoritmo após a iteração M th ou quando o critério BIC começa aumentar.

A Tabela 2 exhibe um resumo dos modelos de aprendizado de máquina utilizados para predição da arrecadação fiscal brasileira.

Tabela 2 – Modelos de Previsão

Modelo	Descrição	Categoria de aprendizagem
Ridge	Regressão Ridge	Shrinkage
Lasso	Método LASSO	Shrinkage
LassoAdap	Método LASSO Adaptativo	Shrinkage
ElasticNet	Método Elastic Net	Shrinkage
ElasticNetAdap	Método Elastic Net Adaptativo	Shrinkage
Bagging	Método de ponderação Bagging	Ponderação
CSR	Regressões de Subconjunto Completadas	Ponderação
Target	Modelo de fatores Target	Fator
Boosting	Modelo de fatores Boosting	Fator

Fonte: Elaboração dos autores.

3.5 Acurácia dos modelos

O desempenho dos modelos será mensurado com base na Raiz do Erro Quadrático Médio (RMSE), fora da amostra. O RMSE é expressado pela Equação (10):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2} \quad (10)$$

em que y' denota o valor predito, y denota o valor efetivo e n representa o número de observações.

4 Resultados e discussão

A periodicidade dos dados utilizados para previsão da arrecadação tributária varia de janeiro de 2002 a outubro de 2021, com frequência mensal. A Tabela 3 mostra as previsões da arrecadação para diferentes horizontes. Nesse caso, as previsões são para oito períodos à frente, para um período acumulado de quatro meses à frente e, por fim, para um período acumulado de oito meses à frente. A amostra de previsão ou de teste corresponde a 20% da amostra completa, ou seja, 80% da amostra foi destinada ao treino na estimação. Estratégia de previsão baseada em [Medeiros et al. \(2021\)](#).

Além das previsões pelos métodos de aprendizado estatístico, já detalhados na Seção 3, também realizamos dois tipos de previsão alternativos: o primeiro é realizado a partir da média das previsões de todos os modelos de aprendizado estatístico, para cada horizonte de previsão considerado, incluindo os horizontes cumulativos; por sua vez, o segundo tipo de previsão considerado é através da mediana de todos os modelos.

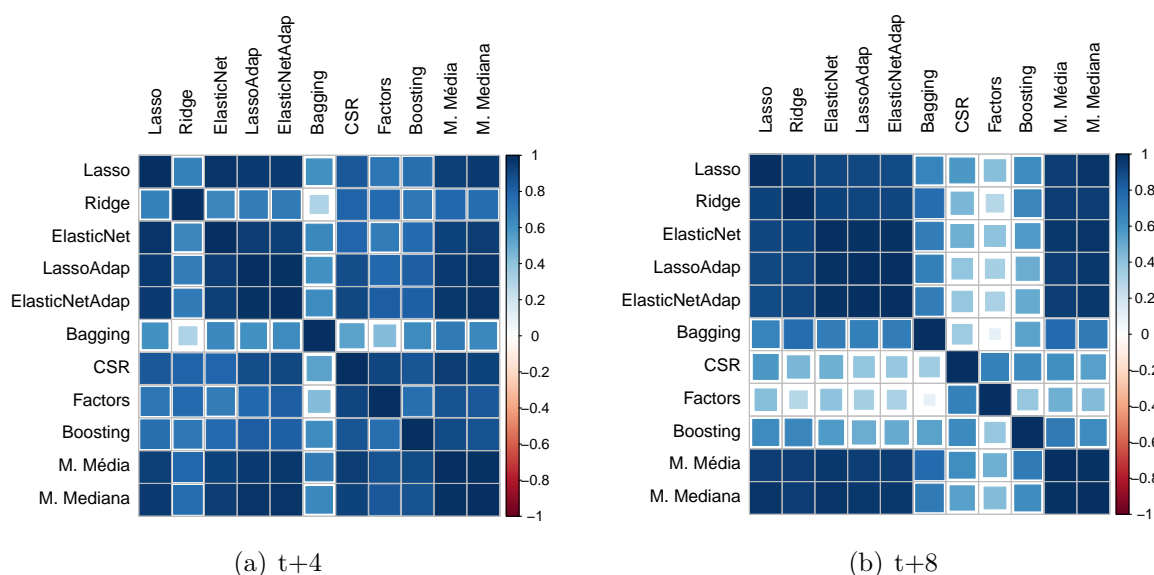
Tabela 3 – RMSE para fora da amostra - Previsão de arrecadação - Novembro/2017 a Outubro/2021

Modelo	Horizonte									
	1	2	3	4	5	6	7	8	4m	8m
Lasso	0.1393	0.1400	0.1350	0.1444	0.1592	0.1744	0.1694	0.1903	0.1099	0.1147
Ridge	0.1264	0.1274	0.1536	0.1683	0.1743	0.1812	0.1561	0.1523	0.1281	0.1251
ElasticNet	0.1089	0.1176	0.1241	0.1291	0.1584	0.1514	0.1560	0.1753	0.1060	0.1106
Target	0.1433	0.1826	0.1717	0.1921	0.1915	0.1928	0.1955	0.1963	0.1461	0.1533
Baggins	0.1749	0.1817	0.1734	0.1484	0.1984	0.1963	0.1957	0.2162	0.1302	0.1262
Boosting	0.1448	0.1472	0.1535	0.1498	0.1583	0.1564	0.1623	0.1587	0.1382	0.1358
ElasticNetAdap	0.1264	0.1354	0.1347	0.1465	0.1600	0.1632	0.1622	0.1914	0.1182	0.1188
LassoAdap	0.1393	0.1400	0.1350	0.1444	0.1592	0.1744	0.1694	0.1903	0.1186	0.1189
CSR	0.1161	0.1364	0.1410	0.1474	0.1521	0.1486	0.1427	0.1414	0.1187	0.1203
M.Média	0.1127	0.1213	0.1277	0.1347	0.1486	0.1506	0.1398	0.1482	0.1110	0.1162
M.Mediana	0.1155	0.1236	0.1319	0.1383	0.1522	0.1498	0.1499	0.1573	0.1137	0.1166

Fonte: Elaboração dos autores.

[1] Nota: Para essa previsão, utilizamos 80% da amostra para treino e 20% da amostra para teste, ou seja, o período usado para previsão, novembro de 2017 a outubro de 2021.

Figura 2 – Correlação das previsões para Arrecadação de novembro de 2017 a outubro de 2021



Fonte: Elaboração dos autores.

[2] Nota: A figura (a) indica as correlações entre as previsões de cada modelo para quarto período à frente, ou seja, o quarto mês. A figura (b) indica as correlações entre as previsões para o oitavo período à frente.

[3] Nota: Essas correlações foram construídas considerando 80% da amostra de treino e 20% para amostra de teste.

De acordo com a Tabela 3, o modelo Elastic Net possui a melhor a acurácia comparado aos demais modelos de aprendizado, para os quatro primeiros horizontes de previsão, com destaque também para o modelo CSR, com alta acurácia para as janelas de previsão iniciais. Garcia, Medeiros e Vasconcelos (2017) destacam o forte poder de previsão do método de ponderação CSR, no caso para variável macroeconômica brasileira. Para os períodos cumulativos, o modelo Elastic Net também supera os demais modelos.

Com relação aos modelos alternativos, constatamos que o modelo médio supera o modelo que leva em consideração sua mediana, para todos os horizontes de previsão, sendo muito próximos para o período cumulativo de oito períodos à frente.

Tabela 4 – RMSE para fora da amostra - Previsão de arrecadação - Dezembro/2015 a Outubro/2021

Modelo	Horizonte									
	1	2	3	4	5	6	7	8	4m	8m
Lasso	0.1141	0.1153	0.1276	0.1343	0.1430	0.1425	0.1423	0.1366	0.1093	0.1044
Ridge	0.1774	0.2701	0.2707	0.2699	0.2698	0.2689	0.2586	0.2554	0.2369	0.2481
ElasticNet	0.1129	0.1162	0.1250	0.1328	0.1399	0.1422	0.1383	0.1415	0.1094	0.1045
Factors	0.1403	0.1556	0.1624	0.1767	0.1758	0.1836	0.1733	0.1702	0.1359	0.1422
Bagging	0.1561	0.1872	0.1517	0.2692	0.1915	0.1769	0.6494	0.1543	0.1368	0.1452
Boosting	0.1369	0.1335	0.1359	0.1456	0.1436	0.1444	0.1496	0.1412	0.1274	0.1195
ElasticNetAdap	0.1257	0.1335	0.1385	0.1432	0.1457	0.1422	0.1401	0.1435	0.1215	0.1134
LassoAdap	0.1297	0.1258	0.1397	0.1479	0.1462	0.1471	0.1461	0.1467	0.1221	0.1143
CSR	0.1149	0.1408	0.1323	0.1448	0.1452	0.1336	0.1327	0.1292	0.1214	0.1166
M.Média	0.1090	0.1188	0.1275	0.1395	0.1385	0.1364	0.1427	0.1239	0.1142	0.1130
M.Mediana	0.1095	0.1152	0.1279	0.1359	0.1404	0.1375	0.1312	0.1307	0.1116	0.1090

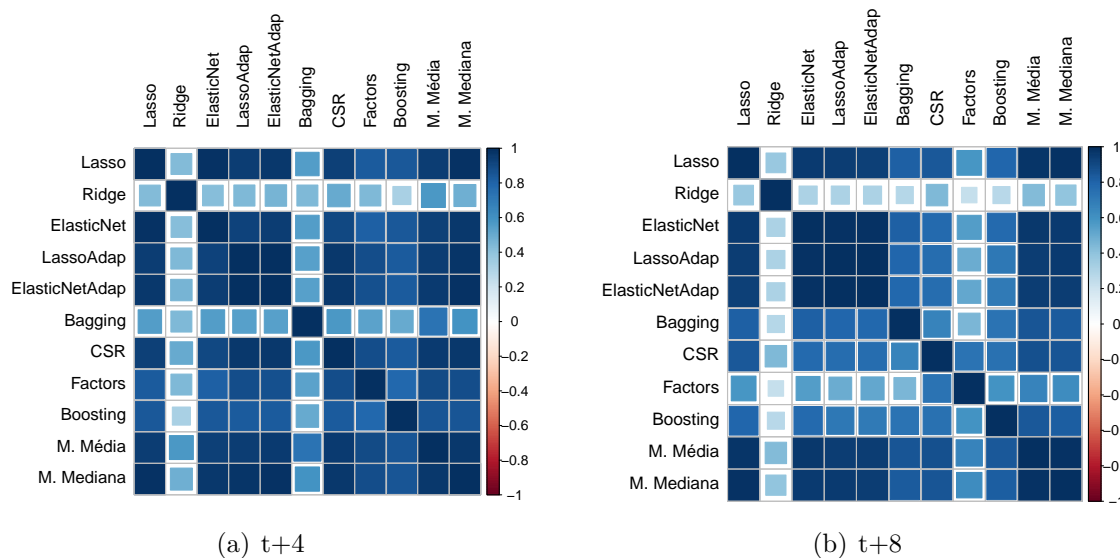
Fonte: Elaboração dos autores.

[4] Nota: Para essa previsão, utilizamos 70% da amostra para treino e 30% da amostra para teste, ou seja, o período usado para previsão, dezembro de 2015 a outubro de 2021.

A Figura 2 mostra as correlações entre as raízes do erro quadrático médio das previsões dos modelos utilizados, para quatro e oito períodos à frente. Conforme esperado, constatamos que todas as correlações são positivas, isso reforça a capacidade de previsão semelhante entre os modelos, apesar de categorias de aprendizagem distintas (shrinkage, ponderação e fator). Também observamos forte correlação do melhor modelo Elastic Net com outros destaques, como CSR, M. Média e M. Mediana. O modelo com menor correlação com os demais é o Bagging, o que reforça a baixa performance daquele modelo de ponderação mostrada na Tabela 3. Uma possível explicação para esse resultado ruim do Bagging é sua baixa capacidade de predição para um conjunto de variáveis e número de observações não tão grande, baseada em sua metodologia de estimação.

Para realização de um período mais longo de previsão com dados mensais, utilizamos 30% da amostra para teste e diminuímos a amostra de treino para 70%, estratégia semelhante a utilizada por Garcia, Medeiros e Vasconcelos (2017). Conforme mostra a Tabela 4, observamos que o modelo Elastic Net manteve sua alta performance de previsão, comparado aos demais modelos, para os períodos iniciais de previsão. Entretanto, os modelos M. Média e M. Mediana foram aqueles com maior acurácia para um período à frente. Para as previsões acumuladas, observamos que o modelo LASSO é altamente competitivo.

Figura 3 – Correlação das previsões para Arrecadação de dezembro de 2015 a outubro de 2021



Fonte: Elaboração dos autores.

[5] Nota: A figura (a) indica as correlações entre as previsões de cada modelo para quarto período à frente, ou seja, o quarto mês. A figura (b) indica as correlações entre as previsões para o oitavo período à frente.

[6] Nota: Essas correlações foram construídas considerando 70% da amostra de treino e 30% para amostra de teste.

A Figura 3 apresenta as correlações das previsões dos modelos de aprendizado estatístico e dos modelos alternativos, M. Média e M. Mediana. Como esperado, todas as correlações são positivas, e a maioria com valores próximos a um. De acordo com a figura (a), o modelo Bagging apresenta maiores correlações com os demais modelos, comparado ao período de previsão da Figura 2. Com relação a figura (b), percebemos que as correlações das previsões do modelo Ridge com os demais são próximas de zero.

5 Teste de robustez

Esta seção apresenta resultados adicionais, baseados em um teste de robustez para a previsão da arrecadação tributária do Brasil. Para esse caso, utilizamos 20% da amostra como teste e 80% da amostra para treino. O exercício de robustez consiste na mudança de frequência de todas as variáveis utilizadas no estudo, inclusive a variável de interesse, arrecadação. Para isso, como os dados originais são mensais, acumulamos os dados em frequência trimestral. Com isso, reduzimos nossa amostra completa de 238 observações para 79 observações. Optamos por excluir a última observação mensal, mês de outubro de 2021, já que essa faz parte do quarto trimestre daquele ano, algo que não é possível para esse estudo.

Tabela 5 – RMSE para fora da amostra - Previsão de arrecadação para fora da amostra - Primeiro trimestre de 2018 ao terceiro trimestre de 2021

Modelo	Horizonte									
	1	2	3	4	5	6	7	8	4m	8m
Lasso	0.1481	0.1216	0.1430	0.1289	0.1289	0.1193	0.1241	0.1292	0.1325	0.1258
Ridge	0.1247	0.1372	0.1418	0.1533	0.1625	0.1688	0.1779	0.1837	0.1372	0.1537
ElasticNet	0.1201	0.1249	0.1188	0.1287	0.1283	0.1260	0.1291	0.1325	0.1198	0.1210
Factors	0.1227	0.2473	0.1920	0.1563	0.1223	0.1308	0.1419	0.1270	0.1581	0.1312
Boosting	0.1577	0.1553	0.1442	0.1440	0.1493	0.1625	0.1748	0.1553	0.1392	0.1349
ElasticNetAdap	0.1419	0.1363	0.1283	0.1476	0.1288	0.1220	0.1229	0.1311	0.1360	0.1270
LassoAdap	0.1579	0.1594	0.1604	0.1624	0.1285	0.1189	0.1254	0.1200	0.1599	0.1358
CSR	0.1519	0.1628	0.1538	0.1588	0.1249	0.1223	0.1264	0.1251	0.1539	0.1352
M.Média	0.1230	0.1407	0.1374	0.1374	0.1226	0.1208	0.1255	0.1232	0.1323	0.1254
M. Mediana	0.1224	0.1315	0.1359	0.1380	0.1257	0.1181	0.1240	0.1256	0.1293	0.1244

Fonte: Elaboração própria dos autores.

[7] Nota: Para essa previsão, utilizamos 80% da amostra para treino e 20% da amostra para teste, ou seja, o período usado para previsão, primeiro trimestre de 2018 ao terceiro trimestre de 2021.

A Tabela 5 apresenta os RMSEs das previsões de arrecadação, com dados trimestrais, para oito períodos à frente, acumulado de quatro períodos à frente (nesse caso, quatro trimestres ou um ano) e acumulado de oito períodos à frente (nesse caso, oito trimestres ou dois anos). Constatamos que o modelo Elastic Net possui a maior acurácia de previsão para os horizontes de previsão iniciais, que supera os modelos M. Média e M. Mediana, que corrobora os resultados para dados mensais apresentados na Tabela 3. Esse resultado justifica a alta performance do modelo de aprendizado Shrinkage, Elastic Net, para previsão da variável fiscal brasileira analisada.

6 Conclusões

Dada a relevância de antecipação e controle da arrecadação tributária brasileira, ainda mais após o período da pandemia de Covid-19, a busca por modelos de previsão de alta acurácia é essencial para tomada de decisão com responsabilidade fiscal, por parte de gestores públicos. Na literatura empírica, existem inúmeros estudos que testam métodos estatísticos para previsão de variáveis macroeconômicas, entretanto, existe uma carência de trabalhos voltados para previsão fiscal, sobretudo para o Brasil.

Este artigo avalia diversas metodologias de alta dimensão, estabelecida na literatura por técnicas de aprendizado estatístico ou *machine learning*, para previsão da arrecadação brasileira para um longo período. Para esse estudo, foram consideradas três categorias de aprendizagem estatística: shrinkage, ponderação e fator. Além disso, estabelecemos modelos médio e mediano como parâmetros de comparação para os demais modelos, algo jamais abordado na literatura de previsão fiscal.

Nossos resultados sugerem: (i) conforme esperado, todas as previsões dos modelos de aprendizado estatístico, apesar de categorias distintas, possuem correlações positivas e altas; (ii) dentre os modelos de aprendizado estatístico, o modelo Elastic Net supera todos para as previsões mensal com uso de 20% e 30% da amostra para teste, para períodos curtos; (iii) o modelo CSR é altamente competitivo para previsões curtas; (iv) para previsões de períodos acumulados, o modelo LASSO tem alta performance; (v) e, por fim, constatamos que o modelo Bagging se torna limitado com aumento dos horizontes de previsão e, sobretudo, diminuição de frequência das variáveis.

Este estudo contribui para gestores e analistas públicos responsáveis por formulação de políticas e tomadas de decisão, que precisam prever receitas fiscais para manutenção e controle de gastos com responsabilidade fiscal. Nossos resultados motivam pesquisas futuras que considerem mais indicadores fiscais para previsão da arrecadação tributária, que podem ser construídos a partir de análise de sentimento textual.

Referências

- BAI, J. Inferential theory for factor models of large dimensions. *Econometrica*, Wiley Online Library, v. 71, n. 1, p. 135–171, 2003.
- BAI, J.; NG, S. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, Elsevier, v. 146, n. 2, p. 304–317, 2008.
- BAI, J.; NG, S. Boosting diffusion indices. *Journal of Applied Econometrics*, Wiley Online Library, v. 24, n. 4, p. 607–629, 2009.
- BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996.
- CALDEIRA, J. F.; MOURA, G. V.; SANTOS, A. A. Previsões macroeconômicas baseadas em modelos tvp-var: evidências para o brasil. *Revista Brasileira de Economia*, SciELO Brasil, v. 69, p. 407–428, 2015.
- CAMPOS, C. Métodos econométricos aplicados à previsão da arrecadação de receitas federais. *Receita Federal do Brasil. ESAF*, 2009.
- ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions. *Journal of Econometrics*, Elsevier, v. 177, n. 2, p. 357–373, 2013.
- ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, Elsevier, v. 54, p. 86–110, 2015.
- FRANKEL, J. Over-optimism in forecasts by official budget agencies and its implications. *Oxford Review of Economic Policy*, Oxford University Press, v. 27, n. 4, p. 536–562, 2011.
- GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 679–693, 2017.
- GIACOMINI, R.; WHITE, H. Tests of conditional predictive ability. *Econometrica*, Wiley Online Library, v. 74, n. 6, p. 1545–1578, 2006.
- GU, S.; KELLY, B.; XIU, D. Empirical asset pricing via machine learning. *The Review of Financial Studies*, Oxford University Press, v. 33, n. 5, p. 2223–2273, 2020.
- HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

- INOUE, A.; KILIAN, L. How useful is bagging in forecasting economic time series? a case study of us consumer price inflation. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 482, p. 511–522, 2008.
- KAMLET, M. S.; MOWERY, D. C.; SU, T.-T. Whom do you trust? an analysis of executive and congressional economic forecasts. *Journal of Policy Analysis and Management*, Wiley Online Library, v. 6, n. 3, p. 365–384, 1987.
- KLIESEN, K. L.; THORNTON, D. L. et al. How good are the government’s deficit and debt projections and should we care? *Federal Reserve Bank of St. Louis Review*, Citeseer, v. 94, n. 1, p. 21–39, 2012.
- KOCK, A. B.; CALLOT, L. Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, Elsevier, v. 186, n. 2, p. 325–344, 2015.
- KRAUSE, G. A.; DOUGLAS, J. W. Institutional design versus reputational effects on bureaucratic performance: Evidence from us government macroeconomic and fiscal projections. *Journal of Public Administration Research and Theory*, Public Management Research Association, v. 15, n. 2, p. 281–306, 2005.
- MCCRACKEN, M. W.; NG, S. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 34, n. 4, p. 574–589, 2016.
- MEDEIROS, M. C.; MENDES, E. F. 1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, Elsevier, v. 191, n. 1, p. 255–271, 2016.
- MEDEIROS, M. C.; VASCONCELOS, G. F.; VEIGA, Á.; ZILBERMAN, E. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 39, n. 1, p. 98–119, 2021.
- PLESKO, G. A. The accuracy of government forecasts and budget projections. *National Tax Journal*, The University of Chicago Press, v. 41, n. 4, p. 483–501, 1988.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.
- ZOU, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, Taylor & Francis, v. 101, n. 476, p. 1418–1429, 2006.
- ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005.
- ZUANAZZI, P. T.; ZIEGELMANN, F. A. Previsões para o crescimento do pib trimestral brasileiro com séries financeiras e econômicas mensais: uma aplicação de midas. *Economia Aplicada*, SciELO Brasil, v. 18, p. 295–318, 2014.

APÊNDICE A – Estatísticas descritivas das variáveis utilizadas

Variável	Média	Mediana	Máximo	Mínimo	D. Padrão	Observações
Arrecadação	80444.95	79539.15	180221.24	22682.45	37847.86	238
Papelão ondulado	261277.91	269930.00	371179.00	157096.00	47699.40	238
Produção industrial (produtos de borracha e material plástico)	92.21	91.80	112.50	55.10	9.21	238
Produção industrial (celulose, papel e produtos em geral)	95.71	97.35	117.90	67.80	9.77	238
Produção (geral)	91.43	91.45	112.60	60.30	9.92	238
Produção (indústria de transformação)	91.25	91.15	113.70	57.70	10.28	238
Produção (bens de capital)	84.19	80.00	127.10	37.70	19.14	238
Produção (bens intermediários)	93.03	93.55	111.40	66.80	8.74	238
Produção (bens de consumo)	90.90	91.30	116.30	53.50	11.00	238
Produção total de autoveículos	223.69	219.15	352.33	1.85	58.65	238
Vendas de autoveículos (total)	200849.90	213283.00	383535.00	1323.00	82889.75	238
Consumo de energia elétrica (comercial)	6023.47	6254.50	8204.00	3450.00	1389.92	238
Consumo de energia elétrica (industrial)	13796.83	14075.00	15886.00	11.00	1858.12	238
Consumo de energia elétrica (total)	34622.08	36344.00	43416.00	22324.00	5359.67	238
Volume vendas no varejo (automóveis, motocicletas, partes e peças)	77.82	80.00	133.70	37.70	21.90	238
IGP-M	0.69	0.57	5.19	-1.10	0.92	238
Índice Geral de Preços - disponibilidade interna (IPA-M)	0.69	0.53	5.84	-1.24	0.92	238
Índice de Preços por Atacado - Mercado (IGP-M)	0.77	0.63	6.73	-1.77	1.28	238
INPC	0.52	0.44	3.39	-0.30	0.45	238
IPCA	0.48	0.44	2.53	-0.35	0.38	238
Produção de caminhões	10164.34	9384.50	22647.00	403.00	4099.02	238
Índice Volume de Vendas no Varejo	79.02	85.45	130.30	42.80	20.35	238
Contribuição - Finsocial / Cofins	13216.69	13435.90	26314.11	3495.75	5731.81	238
Imposto sobre a Renda - Pessoas Físicas	1876.11	1397.53	11663.13	156.30	1749.50	238
Imposto sobre a Renda - Pessoas Jurídicas	8689.31	6840.00	37839.48	1629.19	5942.01	238
Dívida Líquida do Setor Público - Interna	49.20	45.78	75.91	36.83	9.99	238
Dívida Líquida do Setor Público - Total	44.92	45.98	62.54	30.01	9.09	238
Dívida Líquida do Setor Público - Externa	-4.29	-6.76	17.20	-15.82	7.97	238
Dívida Líquida do Governo Geral	45.59	45.32	67.16	30.71	9.61	238
Dívida Fiscal Líquida	41.57	36.66	74.48	29.50	12.47	238
NFSP sem desvalorização cambial	4.04	2.85	13.75	0.00	3.03	238
Dívida mobiliária	2968422.49	2508667.55	7232573.36	699812.73	1842412.33	238
Saldo da carteira de crédito	1999196.59	2010899.00	4511956.00	337735.00	1220041.75	238
Selic	0.92	0.89	2.08	0.13	0.40	238
Taxa de câmbio	2.85	2.51	5.65	1.56	1.05	238

Fonte: Elaboração dos autores.